KNN-防御：使用最近邻搜索抵御三维对抗性点云

Nima Jamali[1], Matina Mahdizadeh Sani[1], Hanieh Naderi[2*], Shohreh Kasaei[3*]

[1]School of Computer Science, University of Waterloo.
[2]Department of Data Science and Technology, School of Intelligent Systems Engineering, University of Tehran.
[3]Department of Computer Engineering, Sharif University of Technology.

*Corresponding author(s). E-mail(s): hanieh.naderi@ut.ac.ir; kasaei@sharif.edu;
Contributing authors: nima.jamali@uwaterloo.ca; m3mahdiz@uwaterloo.ca;

Abstract

Deep neural networks (DNNs) have demonstrated remarkable performance in analyzing 3D point cloud data. However, their vulnerability to adversarial attacks —such as point dropping, shifting, and adding—poses a critical challenge to the reliability of 3D vision systems. These attacks can compromise the semantic and structural integrity of point clouds, rendering many existing defense mechanisms ineffective. To address this issue, a defense strategy named KNN-Defense is proposed, grounded in the manifold assumption and nearest-neighbor search in feature space. Instead of reconstructing surface geometry or enforcing uniform point distributions, the method restores perturbed inputs by leveraging the semantic similarity of neighboring samples from the training set. KNN-Defense is lightweight and computationally efficient, enabling fast inference and making it suitable for real-time and practical applications. Empirical results on the ModelNet40 dataset demonstrated that KNN-Defense significantly improves robustness across various attack types. In particular, under point-dropping attacks—where many existing methods underperform due to the targeted removal of critical points—the proposed method achieves accuracy gains of 20.1 % , 3.6 % , 3.44 % , and 7.74 % on PointNet, PointNet++, DGCNN, and PCT, respectively. These findings suggest that KNN-Defense offers a scalable and effective solution for enhancing the adversarial resilience of 3D point cloud classifiers. (An open-source implementation of the method, including code and data, is available at https://github.com/nimajam41/3d-knn-defense).

Keywords: 3d Point Clouds, Adversarial Defense, Adversarial Attack, Manifold Assumption

1 介绍

深度神经网络（DNNs）在广泛的机器学习任务中取得了显著的成功，尤其是在图像分类 [1–3] 和图像分割 [4–6] 领域。尽管有这些成功，它们仍然对对抗样本 [7] 高度脆弱。这些样本由经过微妙修改的输入组成，以一种欺骗模型做出错误预测的方式更改，而输入本身没有明显变化。

近年来，已经提出了若干 2D 对抗样本生成算法 [8–16] 。这些算法通常可以根据攻击者的访问级别和所使用的技术进行分类。从这些角度来看，对抗攻击可以分为不同的组，包括白盒攻击和黑盒攻击、基于梯度的攻击和基于优化的攻击，以及目标攻击和非目标攻击。

在白盒攻击场景中，对手可以完全访问目标模型的架构、参数和梯度。基于梯度的攻击，如 FGSM (快速梯度符号法) [8] 、JSMA (雅可比显著性图攻击) [10] 和 PGD (投影梯度下降) [11] ，利用模型的梯度来生成对抗性示例。基于优化的攻击，例如 DeepFool [12] 和 C & W [13] ，尝试最小化特定的目标函数以创建对抗性示例。基于决策的攻击，例如 DBA [14] ，直接专注于沿受害模型的决策边界进行优化。所有这些方法都被认为是白盒攻击。相比之下，在黑盒攻击场景中，攻击者对目标模型的架构或参数的访问有限或没有访问权限，必须采取替代方法来生成对抗性示例 [15, 16] 。白盒和黑盒攻击都可以进一步分类为有目标和无目标攻击。在有目标攻击中，攻击者旨在欺骗模型将对抗性示例分类为某个特定的目标类别。相比之下，无目标攻击试图导致误分类而不针对特定类别。

LiDAR 传感器因其能够生成周围物体的高精度和高密度 3D 表示的能力，最近成为一个非常热门的话题。它们广泛应用于各个领域，如自动驾驶 [17, 18] 和机器人技术 [19, 20] 。随着其受欢迎程度的不断提高，研究人员愈加专注于研究点云，这是这些传感器的直接输出。点云是没有特定顺序的 3D 点集，用于表示真实物体的形状。虽然在许多应用中是非常重要的资源，但由于其不规则的数据格式，使得使用传统 DNNs 处理它们时面临主要挑战。

为了解决这个问题，PointNet [21] 提出了一种能够处理无序集合架构，同时对不同变换保持不变。自从 PointNet 被引入以来，3D 点云分类取得了诸多进展。研究人员开发了多种架构，如 PointNet++ [22] 、DGCNN [23] 、PCT [24] 和 CurveNet [25] ，这些架构在分类任务中实现了更高的准确性。

最近的研究表明，点云深度神经网络同样容易受到对抗性攻击的影响 [26–31] 。与二维攻击类似，三维点云攻击可以使用相同的分类方案进行广泛分类，包括白盒和黑盒攻击、基于梯度和基于优化的攻击，以及有针对性和无针对性的攻击。然而，使得三维攻击独特的是基于点云内在属性的新型攻击的出现。在三维点云攻击的背景下，由于其集合结构，对三维点云的攻击可以分为三类：点移动攻击，通过移动现有点的坐标来生成新样本；点生成攻击，通过向集合中添加新点来诱导错误分类；以及点丢失攻击，通过从集合中移除某些点来创建对抗性示例。其中一些攻击能够生成可构造的对抗性示例 [32, 33] ，这对许多对安全性要求高的三维点云应用构成了威胁。

为了使这些网络能够抵御三维对抗性点云，近年来已经提出了几种方法 [34–37] 。其中一种方法 [34] 使用原始数据与对抗性例子的组合来训练模型。另一种方法是由 DUP-Net [35] 提出的，它利用统计异常值移除（SOR）来去除被对抗攻击添加或移动的异常点，然后将清理后的点云数据输入到一个上采样网络中 [38] 。IF-defense [36] 是另一种方法，它使用深度隐式函数从对抗性样本中重建原始样本。这些方法主要关注表面重建 [34–36] 和保持均匀分布 [35, 36] ，这可能对抗删除攻击来说不是最理想的。所提出的 KNN-Defense 通过优先保留点云流形结构来解决这一局限性，从而对各种类型的攻击，特别是删除攻击，提供更有效的防御。

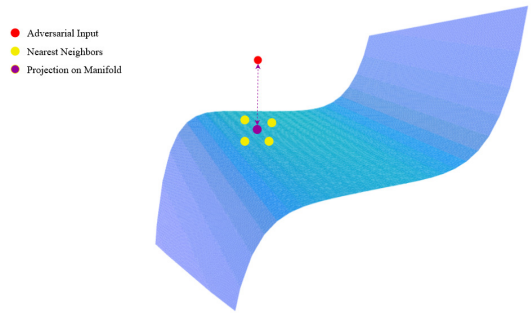


Fig. 1: 在 KNN 防御中的流形假设。干净的点云被认为位于一个 n 维流形上，而对抗样本则偏离这个流形。所提出的方法通过最近邻将对抗输入投影回流形上，以进行稳健的分类。

这项工作介绍了一种防御算法，KNN-Defense，它利用了特征空间中的流形假设和语义邻域一致性。在这一假设下，干净的 3D 点云位于低维流形上，而对抗性的例子则偏离了流形。

如图 1 所示，该方法通过使用与训练集特征空间相似性将扰动样本投射回数据流形来恢复扰动样本。由于流形并不是显式已知的，该方法利用从训练数据中提取的特征对流形进行逼近，并通过聚合其最近邻的 softmax 输出来分类每个输入。

为了提高推理的可靠性，我们采用了三种加权函数——均匀、基于熵和基于多样性的权重，来根据预测置信度平衡邻居的贡献。

与许多现有方法不同，KNN-Defense 不依赖于手工制作的几何先验、架构修改或重新训练。其轻量化和与架构无关的设计使得快速推理和实际扩展变得可能。实证结果证实，KNN-Defense 在多个架构中提高了对点移动和点删除攻击的鲁棒性，确立了其作为 3D 点云分类可扩展解决方案的有效性。

2 相关工作

在本节中，我们提供了 3D 点云和常用距离度量的概述。总结了先前关于 3D 点云分类的研究，并讨论了现有的 3D 点云对抗攻击和防御方法。

2.1 点云及其距离度量

点云是一种用于表示多维数据的数据结构，包括三维对象。具体来说，三维点云由在物体表面采样的点组成。这些点可以表示为向量，向量包含点在 x 、y 和 z 轴上的坐标。例如，一个包含 n 个点的点云可以定义为 $P = \{p_i | i = 1, 2, \ldots, n\}$ ，其中 p_i 表示点云中的第 i 个点。除了几何坐标之外，像颜色或强度这样的外部信息也可以保存在点云中。

由于点云是无序的点集合，传统用于二维图像的距离度量无法直接适用于点云，除非两个点云之间建立了一对一的映射。计算点云之间距离的两种最流行的方法是 Hausdorff 距离和 Chamfer 距离。Hausdorff 距离测量一个点云中某一点到另一个点云中其最近点的最大距离，而 Chamfer 距离则使用求和函数代替最大值。

2.2 用于点云分类的深度学习模型

PointNet [21] 是第一个直接应用于三维点云的深度网络。它使用多层感知器和最大池化（作为对称函数）来提取每个点云的全局特征。相比之下，PointNet++ [22] 利用 PointNet 在层次结构中考察分类中的局部特征。动态图 CNN (DGCNN) [23] 通过构建每个点的 k 最近邻的图并将 EdgeConv 应用于这些图，将准确性进一步提升，从而能够捕捉点云的几何结构以及局部特征。点云变换器 (PCT) [24] 采用变换器架构来处理三维点云数据，通过将点云转换为高维特征空间。使用自注意力机制，它在处理特定点时，根据相似性权衡所有其他点的影响。CurveNet [25] 是另一种架构，通过利用点云内假设曲线的聚合而实现高精度。尽管它们在架构上多样化并表现优良，所有这些网络仍然容易受到对抗性示例的影响，通过几乎觉察不到的变化欺骗模型。

2.3 对抗性攻击方法

基于 Carlini & Wagner （C & W）[13] 优化框架，[26] 提出了通过点偏移（使用 l_2 范数）和点添加（使用 Chamfer 和 Hausdorff 距离）攻击生成对抗性点云的各种方法。KNN 攻击 [32] 结合了 KNN 和 Chamfer 距离度量来生成可构建的对抗样本。此外，众所周知的 2D 梯度攻击算法，如 FGSM [8] 、JSMA [10] 和 PGD [11] 已被改编用于针对 3D 分类网络 [34] 。[27] 开发了一种通过识别点云的显著性图并去除相应点的点丢弃攻击。AdvPC [28] 和 ShapeAdv [39] 使用自动编码器来生成对抗性点云。LG-GAN [40] 为此引入了一种基于生成对抗网络（GANs）的新算法。此外，AOF [41] 将输入分解为低频和高频成分，并通过操纵低频成分来生成对抗性点云。

2.4 对抗性防御方法

在三维对抗防御领域，对抗训练 [34] 被用于利用对抗样本训练分类模型，以增强其对各种攻击的鲁棒性。在 [42] 中，提出了一种方法，通过向对抗点云中添加高斯噪声，将其从对抗子空间中推出。相同研究中提出的另一种方法，称为 SRS，涉及随机选择并删除点云中的 n 点。相比之下，SOR（统计离群值去除）[35] 方法专注于选择性地去除离群点。基于流形假设运行的 DUP-Net [35] 包括使用 SOR 作为预处理步骤以去除离群点。然后将剩余的点输入到上采样网络中以获得最终分类。[36] 引入了两种方法来恢复干净的点云，利用深度隐函数。在 SOR 预处理步骤之后，第一种方法使用隐函数网络从输入点云中构建网格，随后从该网格中采样若干点。后一种方法通过使用几何感知和分布感知损失函数恢复原始形状。通过对点云的索引进行随机变换，[43] 改变了攻击者会计算的梯度，使对抗样本无效，同时分类器对数据的解释能力不受影响。此外，[44] 提出了 Ada3Diff，它使用概率扩散模型 [13] 从对抗点云中恢复干净数据。它首先估计每个点的扰动以评估失真，然后逐步添加噪声并反转它以恢复原始的干净分布。[45] 引入了 PointCutMix，通过基于地球移动距离 (EMD) 计算的最佳对应点混合点云，改进了模型的鲁棒性和泛化能力。最后，[37] 使用频率分析来增强鲁棒性：它应用球谐函数来捕捉点云的频率，然后使用低通滤波器隔离核心形状成分，并将这些过滤后的点云输入分类器以避免对抗性高频噪声。

在二维对抗防御领域的相关工作是由 Dubey 等人提出的 Web-scale 近邻搜索 [46] 。该方法背后的关键概念是流形假设。受这一方法的启发，本文将近邻原则扩展到三维领域，针对无序和不规则的点云结构进行适配，以实现对抗三维扰动的鲁棒性。

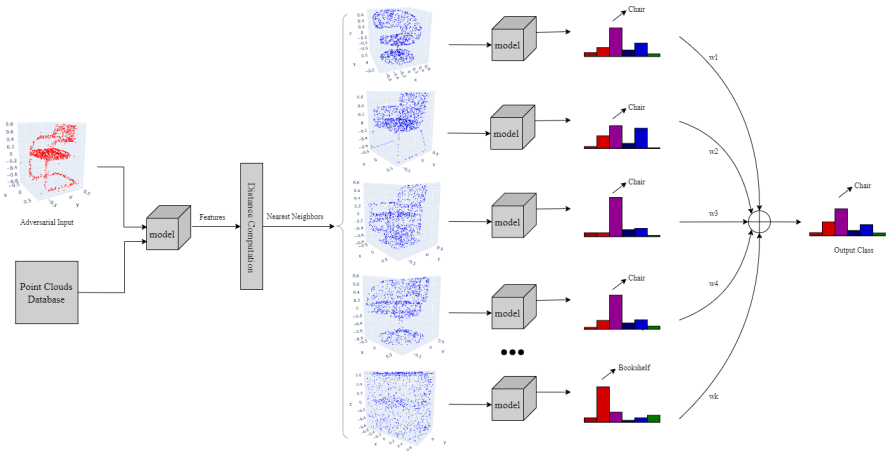


Fig. 2: 提出的 3D KNN-Defense 方法的流程。该方法涉及基于全局特征向量之间的距离在点云数据库中找到给定输入的 k 个最近邻。使用邻居的 softmax 向量的加权平均来预测最终的类别标签。

3 提出的方法

本节提出了一种针对点云分类中对抗样本的 3D 防御方法。该方法基于流形假设，认为干净的点云位于低维流形上，而对抗样本则偏离该结构。主要目标是利用来自训练集的语义相似的邻居，将这些扰动的输入重新投影到数据流形上。为了改进这种投影，探索了三种加权策略。

3.1 防御方法

所提出的防御方法的核心假设是，敌对攻击导致输入点云偏离自然点云流形。因此，目标是识别每个敌对输入在流形上的投影，并对该投影进行分类，而不是敌对输入本身。然而，由于实际的数据流形是未知的，该方法在实际应用中对这一投影进行了近似。

如图 2 所示，输入（无论是对抗性还是干净的）被输入到分类网络，并从特定的隐藏层提取特征。然后计算输入特征与点云数据库中每个样本的特征之间的成对距离，以确定最近的邻居。受 [46] 的启发，该方法使用加权函数计算邻居的 softmax 向量的加权平均。最终的类别预测对应于结果加权平均向量中的最大元素。

算法 1 概述了我们 KNN-Defense 方法涉及的步骤。

3.2 加权函数

可以考虑采用不同的加权方案来计算最近邻的 softmax 向量的加权平均值。类似于 [46] ，所提议的 KNN-Defense 方法中采用了三种距离度量：统一加权（UW）、基于熵的加权（EW）和基于多样性的加权（DW）。

3.2.1 均匀加权（UW）

均匀加权策略为所有 k 个最近邻赋予相同的重要性，而不考虑它们与输入语义上的接近程度。这种方法在计算上简单且稳定，使其适合快速或资源受限的场景。然而，在

Algorithm 1 使用最近邻搜索进行防御

Input: Point Cloud P , outputs of feature layer $\mathcal{L}$ in the classification model $\mathcal{F}$, distance metric $\mathcal{D}$, nearest neighbor number k , weighting function W , training point clouds dataset S_{train}

Output: Final predicted class c

Initialize the nearest-neighbors set $\mathcal{N}(P) = \varnothing$
Compute $\mathcal{L}(P)$ and $\mathcal{L}(P_i')$ for each $P_i' \in S_{train}$
$d_i(P) = \mathcal{D}(P, P_i')$
Sort the values $d_i(P)$ in ascending order, and append the first k neighbors to $\mathcal{N}(P)$
$\forall P_j' \in \mathcal{N}(P)$: Calculate the softmax vector $s_{\mathcal{F}}(P_j')$
$s_{avg} = \sum_{j=1}^{k} W(P_j') s_{\mathcal{F}}(P_j')$
return Prediction class $c = argmax(s_{avg})$

某些情况下，特别是当一些邻居属于语义不一致的类别时，均匀加权可能会降低分类精度。

如图 2 所示，将所有邻居一视同仁可能会导致聚合的 softmax 输出中引入噪声。尽管如此，尽管存在局限性，UW 仍然是一个有效的基线，并对本文提出的整体防御系统的稳健性做出贡献。

3.2.2 基于熵的加权 (EW)

基于熵的加权函数根据样本的 softmax 向量与均匀 softmax 向量（其中所有元素相等）之间的差异来指定权重。softmax 向量 s 的熵通过以下公式计算：

其中 C 表示数据集中类别的数量，s_c 对应 softmax 向量中类别 c 的值。

由于均匀 softmax 向量中所有 s_c 的值都等于 $\frac{1}{C}$ ，因此该向量的熵可以根据公式 (??) 确定为 $\log C$ 。因此，给定 softmax 向量 s 的权重定义为：

$$w = |\log C + \sum_{c=1}^{C} s_c \log s_c|. \tag{1}$$

这种加权策略在过滤不确定的预测方面特别有效，使其非常适合可能存在噪声或模糊邻居的情景。

3.2.3 基于多样性的加权 (DW)

DW（基于多样性的加权）度量评估由 s 表示的 softmax 向量的权重，该评估基于其最大值和次前 M 个值之间的差距。这意味着，softmax 向量的最大元素和其他前 M 个元素之间差距较大的，将被赋予更高的权重。DW 度量采用以下函数来确定按降序排列的软最大向量 $\hat{s}$ 的权重：

$$w = \sum_{m=2}^{M+1} (\hat{s}_1 - \hat{s}_m)^P. \tag{2}$$

在这个方程中，P 和 M 是两个参数，在实验中分别设为 $P = 3$ 和 $M = 20$ 。通过强调具有主导类别预测的 softmax 分布，DW 有助于增强高置信度决策，这在存在类别歧义或对抗性噪声时尤其有价值。

4 实验

在本节中，评估了所提出模型的有效性。结果与现有防御进行比较，以突出所取得的改进。此外，还提供了实验设置的详细说明以及对所提方法相对于现有方法的性能和有效性的深入分析。

4.1 数据集和设置

在实验中，使用了两个数据集：ModelNet40 [47] 和 ScanObjectNN [48] 。ModelNet40 包含 40 个不同对象类别的 12,311 个 CAD 模型。其中的 9,843 个对象用于训练，其余 2,468 个用于测试。使用了该数据集的增强版本，正如在 [26] 中所提供的那样。在这一版本中，从每个对象的表面随机采样 1,024 个点，并归一化以适应于单位球体。另一方面，ScanObjectNN 包含 15 个类别的对象，其中训练集有 11,416 个，测试集有 2,882 个。

本研究中使用的受害者分类模型包括 PointNet [21] 、PointNet++ [22] 、DGCNN [23] 和 PCT [24] 。对于有针对性的攻击，采用了 Shift-L2 [26] 、Add-Chamfer [26] 、Add-Hausdorff [26] 和 Shift-KNN [32] 。对于无针对性的攻击，采用点移除方法 [27] 对干净的点云生成对抗样本。实验中还包括 AdvPC [28] 和 AOF [41] 的有针对性和无针对性版本。在现有的 3D 防御算法中，评估了 SRS [42] 、SOR [35] 、DUP-Net [35] 和 IF-Defense [36] ，以便与提议的 KNN-Defense 方法进行比较。

在 ModelNet40 定向攻击场景中，按照 [26] 中描述的程序，从 ModelNet40 数据集中最大的 10 个类别中随机选择了 25 个样本。对于每个选定的样本，考虑剩余的 9 个类别作为目标标签进行攻击，从而每种定向攻击方法生成了 2,250 个对抗样本。相比之下，非定向攻击应用于整个测试集。

在 KNN-Defense 方法中，特征聚合层的输出被用作点云样本的学习表示。为了确定最佳的 k 参数，进行了若干实验。如图 3 所示，基于 EW 度量，选择的值分别为 PointNet [21] 的 $k = 5$ ，PCT [24] 的 $k = 10$ ，以及 PointNet++ [22] 和 DGCNN [23] 的 $k = 15$ 。

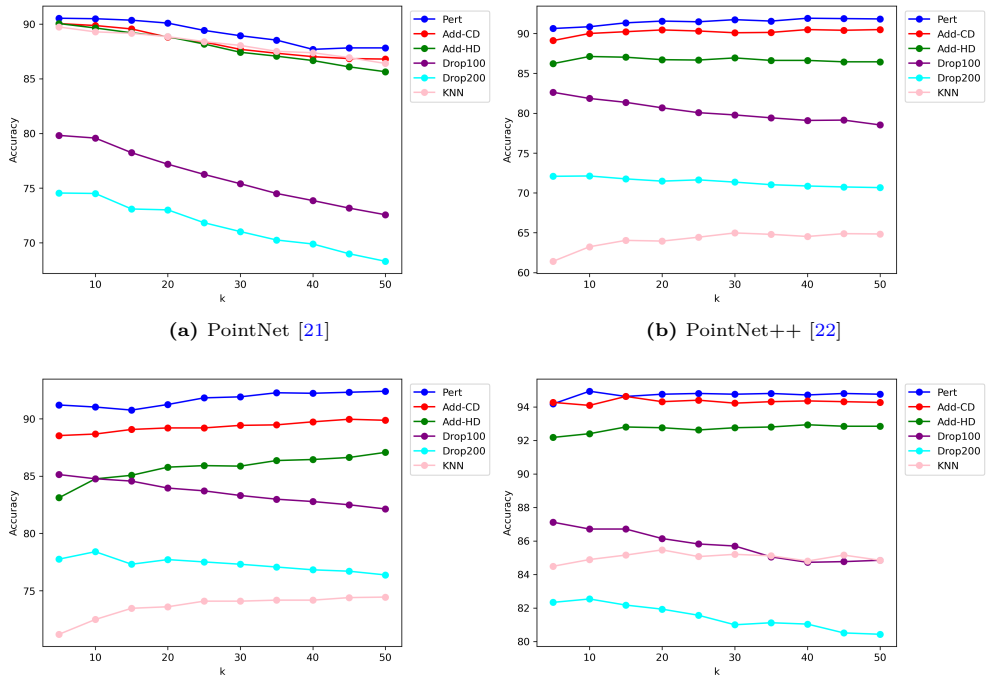
4.2 与最新防御的比较

在本节中，提出的 KNN-Defense 方法的分类准确率与 SRS [42] 、SOR [35] 、DUP-Net [35] 和 IF-Defense [36] 进行比较。

不同防御算法对抗 PointNet [21] 、PointNet++ [22] 、DGCNN [23] 和 PCT [24] 的定向和非定向攻击的分类准确率分别列在表 1 、表 2 、表 3 和表 4 中。此外，使用之前介绍的三种加权函数分别报告了 KNN-Defense 的结果。最高准确率值以粗体显示，次高值以蓝色突出显示。

KNN-Defense 明显优于其他 3D 防御方法，对抗在 ModelNet40 上的无目标点滴攻击 [27] 。它还在对抗 PointNet [21] 和 PCT [22] 的大多数有目标攻击中取得了最先进的结果，同时保持了在不同架构上分类干净图像的合理性能。

表 5 展示了数值结果，证明了提出的防御方法在使用 ScanObjectNN 数据集的 PCT 模型上抵御对抗攻击的有效性。从这张表中可以看出，尽管 KNN-Defense 并不总是优于所有其他防御措施，但它仍然提供了有竞争力的结果，特别是对非定向攻击。这一数据集面临的挑战之一是模型生成高度区分性特征的能力有限，这从基线准确率低于 77 % 可见。这一限制似乎对 KNN-Defense 的性能有一定影响。尽管如此，该方法仍能表现出合理的效果，表明即使在特征表示不理想的情况下仍具有一定的鲁棒性。



(a) PointNet [21] **(b)** PointNet++ [22]

(c) DGCNN [23] **(d)** PCT [24]

Fig. 3: 不同分类模型在面对各种攻击时的准确性，考虑使用 EW 指标作为加权函数的不同 k 值。

Table 1: 不同防御方法在 PointNet [21] 和 ModelNet40 数据集上的对抗攻击准确性比较。

		Targeted Attacks						Untargeted attacks			
Defenses	Clean	Pert	KNN	Add-HD	Add-CD	AdvPC	AOF	Drop100	Drop200	AdvPC	AOF
No defense	88.09	0.00	0.89	0.00	0.00	0.04	0.00	57.09	27.11	1.34	0.00
SRS	76.05	76.27	27.87	75.42	74.36	47.33	9.38	57.78	29.54	26.99	4.78
SOR	76.70	82.71	64.00	83.38	83.20	72.09	38.27	58.14	31.52	48.87	17.18
DUP-Net	77.67	83.24	79.24	84.71	84.44	79.20	61.33	63.29	40.15	63.86	45.70
IF-Defense	84.64	90.22	89.24	90.00	89.42	88.76	79.42	71.88	54.54	82.94	72.73
Ours(UW)	83.55	90.71	89.78	90.04	90.22	88.67	82.89	79.78	74.64	81.36	70.26
Ours(EW)	83.55	90.53	89.73	90.04	90.04	88.76	82.76	79.82	74.55	81.40	70.46
Ours(DW)	83.55	90.62	89.69	89.96	89.82	88.80	82.80	79.82	74.51	81.20	70.38

4.2.1 推理效率

表 6 报告了不同防御方法在所有四种分类架构中的每个点云的平均推理时间。在评估的技术中，SOR 和 IF-Defense 显示出最高的计算成本，平均运行时间分别为 40.32 毫秒和 31.05 毫秒。这些方法依赖于几何操作或内部模型计算，这可能会阻碍在实时系统中的部署。相比之下，所提出的具有均匀加权（UW）的 KNN-Defense 实现了显

Table 2: 在 PointNet++ [22] 和 ModelNet40 数据集上，不同防御方法对抗对抗攻击的准确性比较。

		Targeted Attacks						Untargeted attacks			
Defenses	Clean	Pert	KNN	Add-HD	Add-CD	AdvPC	AOF	Drop100	Drop200	AdvPC	AOF
No defense	89.18	51.20	0.00	41.91	64.62	10.76	3.78	77.55	62.72	16.21	1.50
SRS	81.65	84.40	61.11	78.09	85.82	50.18	23.78	64.91	42.10	50.61	25.45
SOR	83.79	86.31	48.98	88.62	89.64	61.60	34.40	73.70	62.80	52.76	19.61
DUP-Net	80.96	85.82	84.62	85.16	87.60	70.84	48.89	72.81	64.79	59.68	37.28
IF-Defense	82.46	89.24	89.16	88.49	89.24	81.42	69.51	77.19	68.52	75.53	62.68
Ours(UW)	87.03	91.29	63.69	86.93	90.58	68.00	39.78	81.00	72.12	60.66	34.44
Ours(EW)	87.16	91.33	63.51	86.84	90.44	67.96	40.04	81.08	72.04	60.78	34.60
Ours(DW)	87.24	91.16	63.20	86.98	90.27	68.31	41.07	81.20	72.08	60.98	34.48

Table 3: 在对抗攻击下，对不同防御方法在 DGCNN [23] 和 ModelNet40 数据集上的准确性比较。

		Targeted Attacks						Untargeted attacks			
Defenses	Clean	Pert	KNN	Add-HD	Add-CD	AdvPC	AOF	Drop100	Drop200	AdvPC	AOF
No defense	91.86	0.00	0.00	0.00	0.00	20.06	0.00	79.13	63.01	0.00	0.00
SRS	87.07	91.33	73.47	77.56	90.67	61.10	32.98	72.37	56.73	58.93	38.13
SOR	88.49	90.53	17.29	82.22	82.98	52.03	21.19	78.77	68.11	51.69	29.73
DUP-Net	53.81	50.80	19.16	47.16	53.56	30.51	19.85	44.08	35.78	23.73	13.91
IF-Defense	87.32	93.51	90.00	90.36	92.40	80.59	68.23	81.97	73.99	83.96	71.82
Ours(UW)	88.98	90.84	73.38	85.07	89.07	76.13	56.56	84.64	77.43	81.33	68.58
Ours(EW)	89.06	90.76	73.47	85.07	89.07	76.22	56.73	84.56	77.31	81.42	68.84
Ours(DW)	89.10	90.84	73.47	85.07	89.02	76.18	56.93	84.64	77.47	81.82	69.47

Table 4: 不同防御方法在 PCT [24] 和 ModelNet40 数据集上对抗攻击的准确性比较。

		Targeted Attacks						Untargeted attacks			
Defenses	Clean	Pert	KNN	Add-HD	Add-CD	AdvPC	AOF	Drop100	Drop200	AdvPC	AOF
No defense	92.42	72.27	0.00	34.18	69.73	8.49	2.80	80.31	66.00	8.87	2.11
SRS	91.73	92.18	61.02	82.31	91.38	65.24	33.87	83.23	74.59	60.17	21.96
SOR	91.65	93.73	25.02	90.49	93.69	65.6	33.96	82.25	70.95	58.75	19.00
DUP-Net	87.44	92.98	79.11	87.16	92.44	61.11	38.22	76.78	63.94	61.79	38.45
IF-Defense	88.65	92.71	90.89	91.38	92.00	87.07	75.60	83.59	75.49	82.09	68.11
Ours(UW)	90.52	94.76	84.89	92.58	94.36	88.67	81.29	86.75	83.23	80.15	59.85
Ours(EW)	90.48	94.93	84.80	92.53	94.36	88.76	81.38	86.75	83.27	80.19	59.76
Ours(DW)	90.76	94.93	84.80	92.76	94.40	88.89	82.31	86.75	83.10	80.39	60.13

著较低的运行时间，为 5.44 毫秒——比 IF-Defense 快约 6 倍，比 SOR 快超过 7 倍，同时保持强大的对抗鲁棒性。虽然 SRS 是最快的方法，运行时间为 1.57 毫秒，但在早期评估中（表 1–5）显示出较差的防御性能。这些发现突显了 KNN-Defense 在推理速度和鲁棒性之间提供的权衡，使其适合于时间敏感或资源受限的应用。

Table 5: 不同防御方法在 PCT [24] 对 ScanObjectNN [48] 数据集的对抗攻击的准确性比较。

		Targeted Attacks				Untargeted attacks	
Defenses	Clean	Pert	KNN	Add-HD	Add-CD	Drop100	Drop200
No defense	76.66	0.00	0.00	0.00	0.00	63.87	53.30
SRS	73.66	68.32	26.41	44.59	67.14	66.20	60.72
SOR	73.49	67.49	5.34	55.90	68.84	65.58	58.40
DUP-Net	53.78	46.53	38.76	42.16	48.89	48.72	44.27
IF-Defense	59.85	57.18	53.26	54.82	58.19	55.03	50.07
Ours(UW)	70.44	61.00	46.53	44.14	61.07	67.38	62.98
Ours(EW)	70.33	61.00	46.67	44.10	60.83	67.56	63.15
Ours(DW)	70.30	60.96	46.63	44.03	60.79	67.56	63.15

Table 6: 在四个分类模型中，不同防御方法的平均每点云推理运行时间。

Defenses	Time (milliseconds)
SRS	1.57
SOR	40.32
DUP-Net	7.99
If-Defense	31.05
Ours (Uniform)	5.44

5 结论

本文提出了一种针对 3D 点云分类的防御框架，通过利用流形假设和特征空间中的语义邻域关系来提高 3D 模型的鲁棒性。这种方法不是重建表面几何或应用几何先验，而是基于与训练实例的特征空间相似性来恢复被扰动的对抗样本。该方法的一个关键优势是与学习到的特征表示结构的对齐。通过在预训练模型中提取的中间特征上操作，这种方法消除了对架构改变或重新训练的需求。其轻量化设计确保了快速推理，非常适合实时和嵌入式 3D 感知系统。此外，其跨攻击类型的通用性允许有效的防御，无需针对特定威胁模型进行定制。实验评估证实，所提出的方法在现有的 3D 防御方法中表现更优，尤其是在点丢失和点位移攻击下——传统方法经常因关键结构点的选择性移除而失败。这些研究结果强调了在开发可扩展和弹性的 3D 视觉系统防御中基于流形的语义对齐的重要性。

References

[1] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)

[2] Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, pp. 6105–6114 (2019). PMLR

[3] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is

worth 16x16 words: Transformers for image recognition at scale. International conference on learning representations (2021)

[4] Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 801–818 (2018)

[5] Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in Neural Information Processing Systems 34 , 12077–12090 (2021)

[6] Chen, Z., Duan, Y., Wang, W., He, J., Lu, T., Dai, J., Qiao, Y.: Vision transformer adapter for dense predictions. In: The Eleventh International Conference on Learning Representations (2022)

[7] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. In: 2nd International Conference on Learning Representations, ICLR 2014 (2014)

[8] Goodfellow, I., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: International Conference on Learning Representations (2015). http://arxiv.org/abs/1412.6572

[9] Naderi, H., Goli, L., Kasaei, S.: Generating unrestricted adversarial examples via three parameteres. Multimedia Tools and Applications 81 (15), 21919–21938 (2022)

[10] Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z.B., Swami, A.: The limitations of deep learning in adversarial settings. In: 2016 IEEE European Symposium on Security and Privacy (EuroS & P), pp. 372–387 (2016). IEEE

[11] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. International Conference on Learning Representations (2018)

[12] Moosavi-Dezfooli, S.-M., Fawzi, A., Frossard, P.: Deepfool: a simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2574–2582 (2016)

[13] Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 Ieee Symposium on Security and Privacy (sp), pp. 39–57 (2017). Ieee

[14] He, W., Li, B., Song, D.: Decision boundary analysis of adversarial examples. In: International Conference on Learning Representations (2018)

[15] Su, J., Vargas, D.V., Sakurai, K.: One pixel attack for fooling deep neural

networks. IEEE Transactions on Evolutionary Computation 23 (5), 828–841 (2019)

[16] Rahmati, A., Moosavi-Dezfooli, S.-M., Frossard, P., Dai, H.: Geoda: a geometric framework for black-box adversarial attacks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8446–8455 (2020)

[17] Zhou, Y., Tuzel, O.: Voxelnet: End-to-end learning for point cloud based 3d object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4490–4499 (2018)

[18] Shi, S., Wang, X., Li, H.: Pointrcnn: 3d object proposal generation and detection from point cloud. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–779 (2019)

[19] Mei, G., Poiesi, F., Saltori, C., Zhang, J., Ricci, E., Sebe, N.: Overlap-guided gaussian mixture models for point cloud registration. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 4511–4520 (2023)

[20] Qin, Z., Yu, H., Wang, C., Guo, Y., Peng, Y., Xu, K.: Geometric transformer for fast and robust point cloud registration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11143–11152 (2022)

[21] Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 652–660 (2017)

[22] Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Advances in neural information processing systems 30 (2017)

[23] Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M., Solomon, J.M.: Dynamic graph cnn for learning on point clouds. ACM Transactions on Graphics (tog) 38 (5), 1–12 (2019)

[24] Guo, M.-H., Cai, J.-X., Liu, Z.-N., Mu, T.-J., Martin, R.R., Hu, S.-M.: Pct: Point cloud transformer. Computational Visual Media 7 , 187–199 (2021)

[25] Xiang, T., Zhang, C., Song, Y., Yu, J., Cai, W.: Walk in the cloud: Learning curves for point clouds shape analysis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 915–924 (2021)

[26] Xiang, C., Qi, C.R., Li, B.: Generating 3d adversarial point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9136–9144 (2019)

[27] Zheng, T., Chen, C., Yuan, J., Li, B., Ren, K.: Pointcloud saliency maps. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1598–1606 (2019)

[28] Hamdi, A., Rojas, S., Thabet, A., Ghanem, B.: Advpc: Transferable adversarial perturbations on 3d point clouds. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16, pp. 241–257 (2020). Springer

[29] Naderi, H., Dinesh, C., Bajic, I.V., Kasaei, S.: Model-free prediction of adversarial drop points in 3d point clouds. arXiv preprint arXiv:2210.14164 (2022)

[30] Arya, A., Naderi, H., Kasaei, S.: Adversarial attack by limited point cloud surface modifications. In: 2023 6th International Conference on Pattern Recognition and Image Analysis (IPRIA), pp. 1–8 (2023). IEEE

[31] Naderi, H., Bajić, I.V.: Adversarial attacks and defenses on 3d point cloud classification: A survey. IEEE Access 11 , 144274–144295 (2023)

[32] Tsai, T., Yang, K., Ho, T.-Y., Jin, Y.: Robust adversarial objects against deep learning models. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 954–962 (2020)

[33] Cao, Y., Wang, N., Xiao, C., Yang, D., Fang, J., Yang, R., Chen, Q.A., Liu, M., Li, B.: Invisible for both camera and lidar: Security of multi-sensor fusion based perception in autonomous driving under physical-world attacks. In: 2021 IEEE Symposium on Security and Privacy (SP), pp. 176–194 (2021). IEEE

[34] Liu, D., Yu, R., Su, H.: Extending adversarial attacks and defenses to deep 3d point cloud classifiers. In: 2019 IEEE International Conference on Image Processing (ICIP), pp. 2279–2283 (2019). IEEE

[35] Zhou, H., Chen, K., Zhang, W., Fang, H., Zhou, W., Yu, N.: Dup-net: Denoiser and upsampler network for 3d adversarial point clouds defense. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1961–1970 (2019)

[36] Wu, Z., Duan, Y., Wang, H., Fan, Q., Guibas, L.J.: If-defense: 3d adversarial point cloud defense via implicit function based restoration. arXiv preprint arXiv:2010.05272 (2020)

[37] Naderi, H., Noorbakhsh, K., Etemadi, A., Kasaei, S.: Lpf-defense: 3d adversarial defense based on frequency analysis. Plos one 18 (2), 0271388 (2023)

[38] Yu, L., Li, X., Fu, C.-W., Cohen-Or, D., Heng, P.-A.: Pu-net: Point cloud upsampling network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2790–2799 (2018)

[39] Lee, K., Chen, Z., Yan, X., Urtasun, R., Yumer, E.: Shapeadv: Generating shape-aware adversarial 3d point clouds. arXiv preprint arXiv:2005.11626 (2020)

[40] Zhou, H., Chen, D., Liao, J., Chen, K., Dong, X., Liu, K., Zhang, W., Hua, G., Yu, N.: Lg-gan: Label guided adversarial network for flexible targeted attack of point cloud based deep networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10356–10365 (2020)

[41] Liu, B., Zhang, J., Zhu, J.: Boosting 3d adversarial attacks with attacking on frequency. IEEE Access 10 , 50974–50984 (2022)

[42] Yang, J., Zhang, Q., Fang, R., Ni, B., Liu, J., Tian, Q.: Adversarial attack and defense on point sets. arXiv preprint arXiv:1902.10899 (2019)

[43] Zhang, J., Dong, Y., Kuang, M., Liu, B., Ouyang, B., Zhu, J., Wang, H., Meng, Y.: The art of defense: Letting networks fool the attacker. IEEE Transactions on Information Forensics and Security 18 , 3267–3276 (2023)

[44] Zhang, K., Zhou, H., Zhang, J., Huang, Q., Zhang, W., Yu, N.: Ada3diff: Defending against 3d adversarial point clouds via adaptive diffusion. In: Proceedings of the 31st ACM International Conference on Multimedia, pp. 8849–8859 (2023)

[45] Zhang, J., Chen, L., Ouyang, B., Liu, B., Zhu, J., Chen, Y., Meng, Y., Wu, D.: Pointcutmix: Regularization strategy for point cloud classification. Neurocomputing 505 , 58–67 (2022)

[46] Dubey, A., Maaten, L.v.d., Yalniz, Z., Li, Y., Mahajan, D.: Defense against adversarial images using web-scale nearest-neighbor search. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8767–8776 (2019)

[47] Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J.: 3d shapenets: A deep representation for volumetric shapes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1912–1920 (2015)

[48] Uy, M.A., Pham, Q.-H., Hua, B.-S., Nguyen, T., Yeung, S.-K.: Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1588–1597 (2019)